DOCUMENT RESUME

ED 322 749 FL 018 758

AUTHOR Allsion, Desmond; Cheung, Evelyn
TITLE "Good" and "Poor" Writing and Writers: Studying Individual Performance as a Part of Test Validation.
PUB DATE Apr 90
NOTE 20p.; Paper presented at the Annual Meeting of the Regional Language Centre Seminar (Singapore, April 9-12, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cohesion (Written Composition); *English (Second Language); Higher Education; Measurement Objectives; *Student Evaluation; *Test Validity; *Writing Evaluation; *Writing Tests
IDENTIFIERS Writing Tasks

ABSTRACT

A study was conducted to investigate issues of test validation by placing them in a context. The study examined the ways in which instances of "good" and "poor" writing have been identified in the marking of part of an English writing test that is given each year to incoming Arts Faculty students at the University of Hong Kong. The study also investigated the overall performances of "good" and "poor" writers. Implications concern content validity of the test and uses to which test results are put. The study results provided evidence for the following conclusions: (1) the Language Analysis Sessions (LAS Writing Test), the focus of the study, appeared to have validity; (2) the ability to organize and develop a piece of writing that would be considered term paper length is not credited in the test, and therefore, the distinction that the test makes between "good" and "poor" writing is restricted in scope; and (3) the extent of individual learners to write academic text is an issue that the test, at present, fails to address; this discrepancy calls for careful evaluation within the Language Center to provide tests that measure the course's ability to meet the student needs that the courses were designed to do. (GLR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED322749

"Good" and "Poor" Writing and Writers
Studying individual performance as a part of test validation

Desmond Allison and Evelyn Cheung [1]
Language Centre, University of Hong Kong

1. INTRODUCTION

Our paper sets out to investigate issues of test validation by placing them in a context. It examines the ways in which instances of "good" and "poor" writing have been identified in the marking of part of a writing test that is given each year to incoming Arts Faculty students at the University of Hong Kong. It also looks at overall performances of "good" and "poor" writers. Implications concern content validity of the test and uses to which its results are put.

1.1 Assessing writing in English for academic purposes

Our aim in this section is to establish a working perspective for our study. We do not attempt to review the large and growing body of literature on assessment of written English, but confine ourselves to commenting on some familiar questions that are also important for our purposes.

The much-discussed tension between reliability and validity (Davies 1978; cf. Morrow 1981: 13) in language testing is highly evident in the assessment of writing. Let us assume a context in which second language learners are preparing to use English for study purposes, over a wide range of academic subjects. To test the writing abilities of these learners, whether for initial placement purposes or as end-of-course assessment, it is necessary among other things to ensure that (i) a representative range of abilities and tasks is covered; (ii) test marks and grades arise from criteria that are clear and not arbitrary; (iii) information from the test can be interpreted as needed (e.g. for purposes of course evaluation, or for statements about a learner's writing abilities). These issues of principle raise many problems, and have also to be reconciled with practical constraints in the testing situation.

(a) Coverage. The issue of what abilities a writing test is testing calls for both a priori and post hoc validation of test content and of the underlying construct. In a useful discussion (not confined to tests of writing), Weir (1988; 21-25), following

[1] In alphabetical order

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Allison
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

FL 018758

Anastasi (1982; 132), affirms that content validity depends upon the relevance of the individual's test responses to the abilities a test seeks to measure. Simple inspection of test content cannot suffice to establish content validity.(1) Weir also points out that, while statistical data can certainly contribute to our awareness of the construct underlying a test, such data do not come with ready-made conceptual labels. Instead, findings need to be interpreted, in the light of whatever theory we have at our disposal.

The specific abilities or knowledge that a writing test may measure will depend, then, both on the range of tasks set and the ways in which individuals respond to these tasks. We would like immediately to draw attention to work by Cumming (1989) that offers a two-factor account of some second-language writing test data in terms of (i) L2 proficiency and (ii) writing expertise. "Writing expertise" here relates to qualities of discourse organisation, and of content selection and control in response to (complex) demands of the writing task, and is distinct from lexicogrammatical "proficiency". It is relevant to our discussion to ask how far any test of written English in a second language context draws, and should draw, on each of these hypothesised traits.

(b) Marking Criteria. Designing representative writing tasks that can also be reliably graded is clearly a challenging activity. For one thing, concern for reliability cannot be satisfied by eliminating subjectivity from the assessment procedure. Gannon (1985; 39) argues convincingly (in our view) that objective assessment of writing can only be achieved by limiting oneself to trivial or very detailed aspects of writing. We believe Gannon's line of argument holds good in a second-language context, especially if we aim to test writing expertise as well as English language proficiency.

How can subjective marking be made reliable and valid? Current wisdom in this domain (also in oral testing) appears to promote the use of rating scales that draw upon a complex range of criteria yet assign a single mark on the scale. For example, West (1989) presents scales used by two leading examining boards in Britain - the Cambridge Syndicate Marking Scale and Oxford Delegacy Scale. West claims that the use of these scales helps examiners not only to ensure inter-marker reliability on one task but also to "overcome one of the great problems when testing writing : how to compare scripts from different tasks?" On the other hand, White (1985; 18-33) advocates that "holistic scoring" be carried out according to criteria that are developed for each particular essay question and which are interpreted according to the particular group that is tested.(2)

(c) Interpretability. If rating scales are interpreted very differently, then reliability problems may persist. However, high reliability in itself does not guarantee validity. High agreement among raters can still leave us with problems in suggesting just what constructs underlie the written performance - and indeed the

assessments by different raters.

1.2 Assessing writing in one context

We shall pursue these language testing concerns in the context of a group of tests that are administered annually by the Language Centre of the University of Hong Kong to some 400 incoming students of the Faculty of Arts. These tests are given in what are called 'Language Analysis Sessions', and are generally known as "the L.A.S. tests". Separate tests are given for each of the "four skills". (For an outline of the four tests, see Fok 1985; 127-129).

Our focus is on the L.A.S. Writing test, and in fact on one of this test's five sections or 'subtests'. An account of the L.A.S. Writing test and of its rationale, which relates to the demands of a tutorial paper on students' writing abilities, can be found in Low (1982). While various modifications to the test and marking scheme have continued to be made, the basic form of the test remains unchanged.

In brief, the test provides for continuity of topic across a range of tasks. The theme of the paper is on the changes brought about by modernization in a small Chinese village called Ma Lin. The five subtests and tasks are :
1. Proofread (the introduction, which sets out terms and scope of the paper). 2. Write text (from pictures): describe change over time (what has happened to the village). 3. Write text (from note cards); 4. Structure an argument (multiple choice): apply two viewpoints concerning change to sample village. 5. Edit text (a draft conclusion) in light of a reader's comments (re formality and re validity of conclusions).

We shall describe subtest 2 in a little more detail, for reasons that will become clear. In this subtest, the student has to demonstrate his/her linguistic and rhetorical ability to handle the notion of change. Minimal input to the writing task is provided through graphic information showing an agriculturally based village beside the sea. The second picture is rather different, with the area looking more industrialized and populated. In both diagrams, the pictures are labelled to prevent any misinterpretation of what they are supposed to symbolize.

The marking scheme for the L.A.S. Writing test is fully objective for subtest 4 (multiple choice); almost so, for subtest 1 (proofread for error in each line of the text); and subtest 5 (edit parts of the text that have been marked). Alternative possibilities over the years for subtest 1 and 5 have been itemized as acceptable or not. Subtests 2 and 3 involve impression marking, guided by a set of criteria. Four of the five subtests are marked by single markers following the marking scheme; subtest 2 is the exception.

For subtest 2, a total of four markers are involved (in some years six markers): two pairs of markers with each pair marking half the scripts. Markers' meetings before the test allow each pair

to discuss a number of sample scripts that they have marked and to ensure some comparability of standards. Actual test papers are compared in cases of serious discrepancy only, as the timescale for the whole operation is exceedingly tight.

Subtest 2 appears to be the most open-ended. Despite a possible challenge from subtest 3, subtest 2 is in our view the most "writing-like" part of the Writing Test. As we shall see, the subtest has also given rise to some problems during validation studies. We decided to explore this subtest further.

1.3 Validation of the Writing Test

Validation of this test hitherto has been in terms of statistical analysis and interpretation of (a) test results; (b) test-taker attitudes as measured by a post-test questionnaire. Work now planned or in progress aims to supplement and enrich these analyses through (c) more attention to individual test performances (this paper being one contribution here) and (d) interviews with test-takers (planned for 1990).

Without going into great detail, we would like to comment on a few of the findings reported by Lee and Low (1982) with respect to test results.

Lee and Low report a two-factor solution for their test results. They interpret this solution in terms of the relevance of grammatical considerations (but not of extended context) for the two subtests (1 and 3) that load highly on Factor 1, and of the relevance of extended context (but not of a focus on grammar) for the two subtests (4 and 5) that load highly on Factor 2. We have earlier remarked upon the difficulty of assigning conceptual labels to the outcome of a statistical analysis. Let us note here that the relevance of "extended context" to test-taker responses cannot be established solely by pointing to the fact that subtests 4 and 5 occur in the latter part of the test paper.

The omission of subtest 2 from the account is worth noting. Lee and Low report that subtest 2 "had a high error variance (.24) and behaved in a strange way". They therefore decided to omit it from further calculations in their factor analysis. Low (1982: 256) also admits to "some problems with inter-marker reliability" for this subtest, although the reliability coefficient (alpha) of .76 reported by Lee and Low (note 2) appears to us to be respectable given the nature of the task.

We suggest therefore that subtest 2 is an interesting instance of reliability-validity tension. The test appears to be a valid writing task for the testees, but has some reliability problems that limit validity. Since the Lee and Low study, there has been revision of a few items in subtest 4 and 5 and minor changes in the wording of the instructions in subtest 2. The overall rationale of the test however, has remained the same. An attempt to factor analyse the 1986, 1987, 1988 and 1989 results, following the procedures adopted by Low (Principal Component Analysis with varimax rotation), suggests that the test is giving us different

information now. The results will be discussed later in another section.

2. AIMS OF THE STUDY

The overall aim is to extend the study of performance on L.A.S. Writing Subtest 2, and to analyse it in terms of the Writing test as a whole. On the face of it, subtest 2 is a more direct test of writing ability. For this very reason, it may give rise to some of the difficulties incurred by standard essay questions, which include the difficulty of telling "what any given mark actually means" (Low 1982 : 51).

The main aims of the study are to look into individual performance in greater detail, in order to throw more light on the question of what given marks may mean. We are particularly interested in the writers who were identifiable, in terms of their subtest 2 performance, as being well above or well below average. We shall henceforth use "good" and "poor", in double quotes, as labels for two groups of writers that are defined operationally in each year by the subtest scores. (Please see Methods section for details). Whether, to what extent and why these value labels are actually appropriate is what our study is all about.

Two broad questions are: What (possibly quite varied) features of writing could be identified as characteristic of (some or all of) these "good" and "poor" writers? And how did the individual writers, whether "good" or "poor", actually fare in terms of the writing test as a whole?

The choice of lexical, grammatical and discourse features of writing to be examined will be taken up in the Methods section.

A final and very important aim is to evaluate our findings. We shall seek a better understanding of what abilities our chosen subtest actually measures; of whether it does so as clearly and reliably as one would wish; and of whether abilities that are not receiving credit in the marking of the subtest should be given more attention in the marking scheme (either within the subtest or elsewhere in the Writing test). (3)

3. METHODS

(a) Choice of scripts.

We decided to work with selected scripts from two years (1987 and 1989. Plans to include 1988 scripts were set aside because of time restrictions).

As ours is not primarily a statistical study, we are not concerned with equivalence of sample size either within one year or across the two years in the study. Instead, we have chosen to consider as "poor" all scripts receiving the lowest mark on the rating scale (a five point scale) from at least one marker. In practice, this means that the other marker had either also given the lowest mark or a rating one interval above the lowest mark.

Similarly, we define "good" as receiving the highest mark from at least one marker (and, in practice, the highest or highest but one mark from the other marker).

This initially yields figures as follows (with 1988 also included) :

Table 1: Total number of "Good" and "Poor" scripts

	"Good" scripts	"Poor" scripts
1987	8	40
1988	12	24
1989	10	20

It appears generally to be common practice that examiners prove reluctant to use the top band on a scale, even when they are encouraged to use the full range in their marking (North 1990). In the present study, the imbalance between "good" and "poor" for 1987 was so great that we decided first to select scripts which received the lowest mark from both markers (17 scripts) and then select at random another 6 scripts which received the lowest mark from one marker, with the second marker giving the mark which is higher one interval. All "good" scripts for 1987, and all scripts in both categories for 1989, were included in the study.

For reference in this paper, scripts are coded by year (87 or 89); by category (G="Good", P="Poor"); and by an arbitrary number for each individual. Since students do not have access to their subtest grades, confidentiality is preserved by this procedure.

(b) Descriptive categories.

Determining what these categories should be formed an important part of the study. As for testing, so possibly for research itself: objective categories are demonstrable but can be trivial, while more in-depth qualitative assessment of individual writing, however perceptive, will raise problems of verifiability or agreement. Taylor (1986) makes a strong case for detailed study of individual scripts (which he exemplifies) rather than error surveys. For test validation purposes, however, one would need to study quite a number of individuals and still seek some form of generalisability of statement.

We therefore decided to look at the selected scripts using objective measures; to adopt any qualitative assessment where appropriate; and to follow this by a descriptive account of generalizable features, with examples for illustration. For convenience, we list the categories under three headings: lexis, grammar, and discourse.

(c) Objective measures.

Lexis. We decided to do the following: 1) count the number of words (word tokens) as a measure of text length 2) count the number of word types (eg. 'was'; 'are' are counted as one type) 3) calculate type/token ratio as a measure of lexical repetition.

Grammar. The T-unit is adopted as a measure of syntactic complexity (one main clause with any subordinate clauses). We decided to count the number of T-units; the number of error-free T-units; the average length in words of T-units and of error-free T-units. Length of error-free T-units is typically a better predictor of test performance (Larsen-Freeman 1978; Lim 1983).

Discourse. We decided to count the number of paragraphs and note the presence of a recognizable introduction and conclusion.

(d) Qualitative assessments.

We are interested also to assess the quality of the introduction, conclusion, as well as the 'development' of ideas throughout the piece of writing. An impression mark was given using a 3-point scale: 'good';'OK';'poor'. Zero mark was given for the absence of an introduction and conclusion. (By these, we mean some introductory or concluding statements at what we took to be a suitable level of generality).

(e) Statistical correlations.

We are interested in the simple correlation between the five subtests; and between the total writing mark and each of the subtest for 1987 and 1989. Simple correlation was also computed between the two marks for subtest 2. As mentioned earlier, factor analysis was performed on the five subtests for the four years from 1986 to 1989.

It is clear to us that objective and evaluative information needs to be integrated so that useful and meaningful statements can be made. With regard to lexis, type/token ratios have served to draw our attention to individual scripts that have comparatively frequent instances of repetition. Effectiveness of such instances has been evaluated, and one or two cases are reported below. Another one of our descriptive aims is to focus on a small group of lexical items that are related to the notion of change, which is central to subtest 2. We decided to look at how the lexical items "change"; "replace"; "turn"; and "develop" are used in each script. This involves grammatical considerations. We are also interested, at the discoursal level, to look at how connectors are used to link ideas together. We shall report on all this in the Discussion section.

4. FINDINGS

4.1 Quantitative survey of scripts (Table 2; Table 3)

Although the mean scores are fairly crude measures, they tend to support existing research and teacher comments that "good" writers on average would write longer texts than "poor" writers, and would make fewer errors. The difference in the average length of error-free T-units appears to be more pronounced than that in the average length of T-units in the two groups in 1989, but not in 1987. We are aware, of course, that the unequal number of samples between the "good" and "poor" scripts inhibits us from making too strong a claim about trends that have emerged.

The difference in the number of paragraphs appears to be negligible, as both groups tend to use more than a paragraph. It is difficult to interpret the average mark for introduction and conclusion in table 2, because the number of students in 1987 who wrote an introduction and a conclusion is small. However, it is worth noting that more students in the "poor" group (14 out of 23. or 60.9%) wrote a conclusion compared to the "good" group (2 out of 8 or 25%). The average mark for 'development' again seems to show that the "good" group scored higher.

Table 2: Mean scores of 1987 scripts

	L(W)	T/T	TU	W/TU	EFTU	%EF	W/EFT	Pa	Int	Con	Dev
Good	210	.58	15.3	14.1	11.3	73	13.3	2.1	1.0	0.5	2.5
Poor	171	.58	15.3	11.4	5.7	38	10.8	2.0	0.5	0.8	1.4

Table 3: Mean scores of 1989 scripts

	L(W)	T/T	TU	W/TU	EFTU	%EF	W/EFT	Pa	Int	Con	Dev
Good	204	.56	14.4	14.7	11	78	14.7	2.7	1.2	1.0	1.8
Poor	163	.56	12.7	13.0	3.9	32	10.7	2.2	1.0	0.6	1.3

KEY: L(W):length in token words; T/T:token/type ratio; TU:T-unit count; W/TU:average T-unit length in words; EFTU:error-free T-unit count; %EF:percentage of error-free T-units; W/EFT: average length of error-free T-unit; Pa: paragraph count; Int: mark given to introduction; Con:mark given to conclusion; Dev: mark given to development.

4.2 Percentile ranking.

For 1987, 5 of the 8 "good" scripts are the work of the students above the ninetieth percentile. Two scored above the eightieth percentile and the last one near the margin of the eightieth (79.5). As for the "poor" scripts, 7 of the 23 scripts were the work of students scoring below the tenth percentile. Of the remaining 16, one scored above the sixtieth percentile; two more above the fortieth percentile; another three above the thirtieth percentile; three more above the twentieth percentile; and seven more above the tenth percentile.

For 1989, 8 of the 10 "good" scripts on subtest 2 are the work of students scoring above the ninetieth percentile on the Writing test (including the two highest scorers on the test). The other two students were ranked at the 87th and 79th percentile. As for the "poor" scripts, 13 of the 20 scripts were written by students scoring below the tenth percentile (including the two lowest scorers on the test). Of the remaining cases, one scored above the eightieth percentile; a second also scored above the fortieth percentile; a third above the thirtieth percentile; and two more above the twentieth percentile.

It appears that if a student does well on subtest 2, it is highly likely that s/he would perform well on the test as a whole. However, with the "poor" scripts, the picture is less clear. While about 60% of the students in both years rank below the bottom twentieth percentile, the rest could range above the twentieth to the eightieth percentile.

4.3 Simple correlations

Table 4: Overall correlation between mark 1 and mark 2 for subtest 2

1987	.72 (p< .01)
1988	.59 (p< .01)
1989	.63 (p< .01)

Table 5: Correlation among the subtests; and between the writing total and the subtests for 1987

Subtest	1	2	3	4	5	
1		.32	.25	.37	.34	
2			.24	.27	.15	
3				.17	.09	(p < .01)
4					.27	
5						
Total	.70	.62	.63	.62	.57	

Table 6: Correlation among the subtests; and between the writing total and the subtests for 1989

Subtest	1	2	3	4	5	
1		.42	.37	.36	.22	
2			.29	.24	.15	
3				.18	.16	($p < .01$)
4					.32	
5						
Total	.72	.64	.64	.64	.60	

4.4 Factor analysis

Table 7: Results of factor analysis of the writing test (varimax rotation) for 1987

	1987		
	Factor 1	Communality	Percentage of Variance
Subtest 1	.757	.573	40.2
Subtest 2	.624	.389	19.1
Subtest 3	.500	.250	15.3
Subtest 4	.684	.468	13.7
Subtest 5	.574	.329	11.8

Table 8: Results of factor analysis of the writing test (varimax rotation) for 1989

	1989		
	Factor 1	Communality	Percentage of Variance
Subtest 1	.775	.601	42.1
Subtest 2	.662	.439	19.4
Subtest 3	.612	.375	14.9
Subtest 4	.648	.420	13.1
Subtest 5	.519	.270	10.5

Although the results of four years from 1986 to 1989 were factor analysed, only the results for 1987 and 1989 are shown here. In the four years, only one factor could be extracted. (This remained the case when the analysis was repeated after excluding subtest 2,

to ensure exact comparability with the procedure followed by Lee and Low (1982)).

5. DISCUSSION

5.1 Subtest 2 and the L.A.S. Writing Test

The correlations between the two sets of marks for subtest 2 (Table 4) indicate that obtaining high inter-marker reliability remains a problem. Interestingly, this problem persists despite revision of the marking scale. Although "rough justice" is done in that all marks given on the five-point scale correspond to within one band (i.e. the second mark is the same as, or one above, or one below the first mark in any case), it appears that markers are operating with rather different practices that may reflect differences in their constructs of what constitutes "good" or "poor" performance on the task. Further study of this issue would have to examine scripts taken from "the middle" of the distribution.

Results from the factor analysis for the four years suggests that there is a common factor between subtests and all other factors are specific factors. Each subtest therefore can be assumed to measure rather different aspects of language skills. In the four years, subtest 1 is the most predominant. The results here are supported by simple correlations in Table 5 and 6. The writing total correlates positively and significantly, although not very highly, with each of the subtests. However, there are very low correlations among the subtests.

It was mentioned earlier that it was easier to predict a higher total mark for students who performed well in subtest 2, but that the converse does not always hold among the "poor" scripts. What seems to stand out now is that subtest 1, which is the explicit grammatical component of the test is comparatively a better predictor of the overall writing test score. (Note the higher correlation coefficient for total and subtest 1 in both years). Similarly, analysis of the scripts of subtest 2 also suggests that the "good" scripts have comparatively fewer lexico-grammatical errors (Tables 2 and 3).

Part-whole correlations and inter-subtest correlations do not suggest that subtest 2 is markedly unusual (though admittedly this information is not very refined). The tendency for moderate correlations with subtest 1 may reflect an element of "knowledge of the L2/grammatical accuracy" in the marking of subtest 2.

5.2 Measures of individual performance

Comparison of the mean scores for "Good" and "Poor" categories in the same year gives few surprises. With the exception of

type/token ratio (and of identical number of T-units per essay in 1987), means are consistently higher for the "Good" group. The difference in both years is most marked for absolute numbers of error-free T-units (EFTs) per essay. The status of 'average length of EFTs' as a better predictor than 'average length of T-units' for performance on the writing test (cf. Lim 1983: 42) receives support from our 1989 but not our 1987 data.

Type/token ratio is the same for both groups in 1987 and also in 1989. The range, however, is greater for the "Poor" group, from a low of .49 to a high of .70 (two cases) in 1987. The same holds true for 1989, with a range of .44 to .66 (two cases). This reflects two kinds of text : scripts of length comparable to the "Good" group but with rather less lexicogrammatical variety (ie. with more repetition) and hence a low type/token ratio, and very short scripts in which there is less lexical repetition and hence a high type/token ratio. (It is worth noting that grammatical as well as lexical considerations affect outcomes here).

We note that type/token ratio is not an effective measure of expressivity or of explaining why one text is better than another in terms of lexical choice (Carter, 1987:89). However, we find that the type/token data provides an additional reason for examining and comparing certain scripts, and for checking somewhat unusual features of individual scripts within the "poor" and "good" sample.

While there may be a tendency for teachers to approve of lexical variation (suggesting a wider range of available vocabulary), and while this is to some extent reflected in the marking scheme (e.g. ways of signalling "change"), it should be stressed that lexical limitations appear not so much in the amount of repetition as in the uses to which repetition is put. In one "poor" script (87P21), the word "village" is repeated nine times. In this case, the repeated use of the word did not help to clarify the situation being described. Its overuse created a sense of redundancy instead. A "good" script (87G1) however, with a somewhat similar type/token ratio, used the word "village" seven times, (in a comparatively longer text, 261 words compared to 163), and used the word rather more effectively.

> **The walled village which was the people's main dwelling in the past has become a model village mainly for tourist attraction. The village has become a modern town with various economic activities.**

In the three instances here, the lexical repetition provides a framework for a developing argument. Each use of "village" is distinctive. The former "walled village" is now a "model village", and "the village", Ma Lin has become a modern town. The repetition here contributes to coherence.

The listing of the lexemes of "CHANGE" is to some extent arbitrary, (since there are many indirect signals, such as tense shift or lexical contrast that were not looked at) but the procedures are consistent and still informative. The lexemes

change, replace, turn and develop in the context of subtest 2 share this notion of 'current relevance' and of a state of affairs moving gradually from one state to another. Most students used the word change in various grammatical forms in their text and to a lesser extent, replace, turn and develop.

Examination of the occurrences of change as a verb among the "good" scripts in 1987 and 1989 show that in all twelve occurrences (100%), it was used in the present perfect tense. Among the "poor" scripts this tense accounted for 12 occurrences out of a total of 37 (32.4%), and the rest were either in the past simple, present simple, past perfect (once) or present progressive (once). The notion of change in subtest 2 is much more adequately captured using 'dynamic' verbs, which include the few above, in the present perfect form. As far as the word 'change'is concerned, the "good" scripts are consistent in using the present perfect tense, and are probably rewarded for it.

It has long been recognized that improving the range of vocabulary of a second language learner is most effectively done in context, so that the collocational field of the word is also perceived. Our feeling is that while lexical variety is evident in most of the scripts, errors in collocation (and grammar) meant that lexical range per se did not necessarily receive credit from markers.

It is revealing to find that "good" students as well as "poor" students produced a number of anomalous sentences, as in the examples below :

(87G1) ...rice fields have been changed into large and modern factories.
(89G10) The rice fields on the hillside have been changed into factories.
(89G6) ... the fields and gardens have turned into factories.
(89G2) ... the rice fields have been replaced by noisy factories.

Writers of both the "poor" and "good" scripts in the sample do not appear to be sensitive to the inappropriacy of seeing rice fields turn into factories. A possible explanation for this recurring feature may be traced to interference from L1. Direct translation equivalents in Chinese prove to be more acceptable.

With regard to grammar, our findings on average length of T-units also drew attention to unusual scripts in each of our categories. There was one striking instance (89G10) of a "good" script that exhibited quite high control over a very limited grammatical range (no major subordinate clauses). Conversely, some of the "poor" writers attempted more complex grammatical structures that they proved unable to control, as in the following example.

(89P16) **As mentioned above, we will have a case study on Ma Lin, as an example, to show have modenization has changed an traditional agricultural village to an industrial town from 1950s to the present.**

We did not attempt any systematic categorization of grammatical errors within the time available to us. We noticed that form class errors were frequent among the "poor" scripts (e.g. "an industry town"; "agriculture life"; "agricultured"), but were rare among the "good" scripts.

An attempt at describing the discoursal competence of the students was through evaluating the introduction, the conclusion and overall development in the sample texts. We interpreted introduction to mean any 'advance organizer' which gives an overall perspective of change seen from the present or in transition and/or which referred to Ma Lin as an example of the process of change. The instructions regarding subtest 2 are explicit about the purpose of the task, which is to describe physical changes and then socioeconomic changes, but it is left open to the students how they would like to structure the text. In both years only about 50% of the scripts contain introductions. Introductions which merited a '3' are not only found among the "good" scripts, but also among the "poor" ones. In the following example, taken from among the "poor" scripts, the student demonstrates that he can write an introduction, but grammatical inaccuracies and poor development of ideas in the rest of the text possibly affected his final mark.

(89P8) **Ma Lin, an agricultural village in the 1950's. But now, it become a tourist attraction. Let us see how does this village changed in this short period.**

See also the example (89P16) we cited earlier as an instance of attempted grammatical complexity.

Similarly, conclusions are found only in about 50% of the scripts in both years. Compared to introductions, the quality of conclusions is lower. This seems to reflect a general trend among both "good" and "poor" scripts.

Discoursal skills are not only demonstrated through elaborate advance signals at the beginning of the text or through a satisfatory concluding statement. Achieving coherence is attempted through a number of ways. Some students rationalize the changes in Ma Lin through a chain of cause and effect events. Others depend heavily on the use of connectors to move from one subject to another. "Good" scripts generally cover the topic adequately, with very few irrelevant elaborations. There are many cases of misuse of connectors, especially 'moreover' and 'on the other hand' in both "good" and "poor" scripts which warrant further careful study. An example to illustrate this can be found in a "good" script.

(87G2) **In the early 1950's, Ma Lin was only a traditional village. People living there mainly engaged in the primary activities such as fishing since Ma Lin is near to the coastal areas. Besides, with the presence of lowland, so the villagers also engaged in the activities of market gardening, that was, the growing of vegetables and flowers as well as growing paddy ... On the other hand , Ma Lin was a walled village since it locates near the coastal areas, thus the whole village was surrounded by the wall ...**

There seems to be a higher marker tolerance for inadequate signalling inter-sententially, as long as ideas flow reasonably smoothly in the overall comparative description of change. Writers who made errors at this level and who compensated by not making too many grammatical errors elswhere are not penalized.

The demands of this writing task upon discourse awareness and organisational abilities are not in fact particularly elaborate. Some of the writers explicitly organised their material in terms of such dichotomies as "economic" and "social" change, or of "physical" changes and changes in "occupation". Such overt structuring of text was by no means confined to "good" writers in the study. Indeed, some of the "poor" writers were quite sophisticated in this respect.

6. CONCLUSION

What can be learned from all these observations and comments? At this point in our paper, we face the two dangers of offering rash and over-generalised conclusions or of venturing remarks that are so tentative as to be worthless. At the risk of folly, we shall state our conclusions firmly. We do so on the understanding that our investigation of how individual performance was assessed in this test forms part of a wider appraisal of the test's validity, and that the conclusions we propose are inferred from our observations rather than being demonstrated beyond all dispute. In addition, our comments on research methods remain impressionistic.

Our conclusions relate to the content validity of the writing test in the study, the uses to which its results are put, and the implications of the study for test validation as such.

The L.A.S. Writing test provides a fairly wide range of tasks and may appear to achieve a reasonable overall compromise between demands of reliability and those of validity. Our study shows, however, that the grading of individual scripts, on what is ostensibly a subtest concerned with continuous writing, in fact largely reflects the writers' lexicogrammatical proficiency in English. Ability to organise and develop a piece of writing - what Cumming (1989) considers as "writing expertise" - is not noticeably credited in the test. It is also clear that abilities to sustain a complex argument, e.g. in a term paper, are not assessed by the test. The distinction that the test makes between "good" and "poor" writing is consequently restricted in its scope.

The extent to which this limitation is a problem for content validity depends partly on the uses to which results are put, since these uses are relevant to the choice of what to test in the first place. On the one hand, the "poor" writers identified by the test plainly need to develop their proficiency in written English. On the other hand, the courses to which Arts students have been assigned in recent years also place considerable emphasis on helping students to meet the discourse demands of academic communication. Yet discourse abilities or disabilities have not appreciably contributed to the placement of students in terms of these courses. The extent of individual learners' problems with local or global information structure in producing academic text is an issue that the test at present fails to address. At the very least, this discrepancy calls for careful evaluation within the Language Centre, both in terms of what we expect of our tests and of what student needs our courses should principally aim to meet.

Some of the methodological benefits of the study are intangible. Clearly we have examined individual scripts in more detail and more systematically than is normally possible when marking, and certainly when marking test scripts. This has given us a better feel for some of the qualities and limitations in the writing produced under test conditions. Among the practical consequences are that we can select scripts that would repay particular attention in markers' meetings. We can also document the persistence of reliability problems in marking. We can thus

argue for a more sustained approach in future towards achieving greater marking consistency. There needs to be a more rigorous institutional procedure both for realising a marking scheme and for ensuring that interpretations of this scheme converge among individual markers.

The study of individual test scripts has commended itself to colleagues and is being extended to validation of another writing test. We feel the approach could also usefully be applied to studies of how individual markers categorise scripts as particularly "good" or "poor".

The limitations of such work will be apparent. The choice of descriptive categories will depend on the researchers' evaluation of the writing task and doubtless on other considerations such as main research interests. While we make no apology for our own interest in lexis, we recognise that other researchers might have found useful information from a closer look at grammatical categories in our data. The danger here is that we all find what we are looking for, and overlook other potentially useful information. Provided, however, that work of this nature is undertaken to supplement rather than to replace other approaches, we maintain that it offers vital information in the validation of open-ended tests.

NOTES

1. For courses in English for academic purposes, we follow Hughes (1988:42) in basing "content" on course objectives rather than on the syllabus and materials through which a course attempts to meet these objectives. The relevance of learner responses (processes as well as products) will be apparent in such a perspective.

2. White also describes detailed procedures by which meaningful holistic marking can be undertaken successfully in one context. These procedures sound admirable, but are also clearly demanding in terms of time and manpower. We plan to explore their feasibility within the Language Centre.

3. With regard to abilities that may not receive sufficient credit, compare a suggestion by Low (1982:256, note 4) that more attention be given to the transitions made between different parts of the tutorial test paper).

We would like to thank our colleagues, Lily Leung and Jo Lewkowicz for contributing to the selection of descriptive categories used in this study.

REFERENCES

Anastasi, Anne 1982. Psychological Testing. New York and London : Collier Macmillan.

Carter, Ronald 1987. Vocabulary : Applied Linguistic Perspectives. London : Allen and Unwin. (Aspects of English).

Cumming, Alister 1989. 'Writing Expertise and Second Language Proficiency'. Language Learning Vol.39, No. 1; 81-141.

Davies, Alan 1978. 'Language Testing'. Language Teaching and Linguistics Abstracts 11/3 and 11/4.

Fok, C.Y.Y. Angela 1985. 'Language Proficiency and Related Factors'. In New Directions in Language Testing (eds. Lee et al), Oxford: Pergamon Press, pp. 127-135.

Gannon, Peter 1985. Assessing Writing : Principles and Practice of Marking Written English. London : Edward Arnold. (Explorations in Language Study series).

Hughes, Arthur 1988. 'Achievement and proficiency : the missing link?' In Testing English for University Study (ed. A Hughes), ELT Documents 127, Oxford : Modern English Publications (with British Council) pp. 36-42.

Larsen-Freeman, Diane 1978. 'An ESL index of development'. TESOL Quarterly 12/4; pp.439-448.

Lee Yick Pang and Graham Low 1982. 'Towards a Classification of Tests of Language Use : The Dimension of Test Directness'. Journal of Applied Language Study 1/1 (April), pp.17-28.

Lim Ho Peng 1983. 'Using T-unit Measures to Assess Writing Proficiency of University ESL Students'. RELC Journal 14/2, Dec. pp. 35-43.

Low, Graham 1982. 'The Direct Testing of Academic Writing in a Second Language'. System 10/3, pp. 247-257.

Morrow, Keith 1981. 'Communicative language testing : revolution or evolution?' In Alderson, J.C. and A. Hughes (eds). Issues in Language Testing, ELT Documents 111, London : British Council.

North, Brian 1990. 'The Assessment of Writing'. Letter in IATEFL Newsletter 106, Jan., p. 6.

Taylor, Gordon 1986. 'Errors and Explanations'. Applied Linguistics 7/2, pp. 144-166.

Weir, Cyril 1988. Communicative Language Testing. Exeter : University of Exeter.

West, Richard 1989. 'The Testing of Writing'. In Testing (Newsletter of the IATEFL Special Interest Group on Testing), no. 5, November.

White, Edward M. 1985. Teaching and Assessing Writing. San Francisco : Jossey-Bass.